Discussion

Lan Wang[1] and Ben Sherwood[2]

[1] School of Statistics, University of Minnesota, Minneapolis, USA
Email: wangx346@umn.edu
[2] Department of Biostatistics, Johns Hopkins University, Baltimore, USA
Email: bsherwo2@jhu.edu

We congratulate Drs. Yang, Wang and He on a very interesting and insightful paper. Quantile regression has gained considerable popularity in all areas where traditional least squares regression is useful. While least squares regression focuses on modeling the center of the conditional distribution, quantile regression provides valuable information about the whole conditional distribution. However, unlike inference for least squares regression, constructing confidence intervals for quantile regression is known to be highly nontrivial even in standard practice of linear quantile regression with complete data, as the asymptotic covariance matrix involves the unknown error density function. This paper explores a general approach based on asymmetric Laplace working likelihood and Bayesian calculation to make asymptotically valid inference for quantile-regression based statistical models. The usefulness of the proposed approach is convincingly demonstrated on two examples: ordinary quantile regression with complete data and quantile regression with fixed censoring in the paper.

The proposed method is computationally and theoretically attractive. In this discussion, we demonstrate the generality of the proposed approach on a third example: quantile regression with missing covariates. The problem of missing data arises in diverse areas of research. When naively ignoring missing data, biased estimation is often resulted. Sherwood, Wang and Zhou (2013) studied weighted quantile regression for unbiased estimation when the covariates are missing at random. For the ith subject, we observe (Y_i, X_i'), $i = 1, \ldots, n$, where Y_i is the response variable, $X_i = (W_i', V_i')'$ with W_i denoting a p-vector of covariates that is always fully observed and V_i denoting a q-vector of covariates that may contain some missing components. Let R_i be the missing data indicator, $R_i = 1$ if V_i is fully observed, and $R_i = 0$ otherwise. We assume that V_i is missing at random in the sense that $P(R_i = 1 \mid Y_i, X_i) = P(R_i = 1 \mid Y_i, W_i)$. This probability can be modeled using parametric, semiparametric or nonparametric models (referred to as propensity score models). In practice, parametric modeling, such as

logistic regression, is popular which assumes $P(R_i = 1 \mid Y_i, X_i) = \pi(T_i, \gamma)$, where $T_i = (Y_i, W_i')'$, $\pi(\cdot, \gamma)$ has a form that is known up to a finite-dimensional parameter γ.

Assume that the τth conditional quantile of Y_i given X_i is $Q_{Y_i|X_i}(\tau) = X_i'\boldsymbol{\beta}(\tau)$, $0 < \tau < 1$, where $\beta(\tau)$ is a vector of unknown quantile regression coefficients. Sherwood, Wang and Zhou (2013) proposed the inverse probability weighted quantile regression estimator

$$\widehat{\boldsymbol{\beta}}_n^W = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{R_i}{\pi(T_i, \widehat{\gamma})} \rho_\tau(Y_i - X_i'\boldsymbol{\beta}), \tag{1}$$

where $\widehat{\gamma}$ is a suitable estimator of γ. Sherwood, Wang and Zhou (2013) proved that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^W - \boldsymbol{\beta}(\tau)) \to N(0, D_1^{-1}(D_0 - D_2' I(\gamma)^{-1} D_2) D_1^{-1})$ in distribution under regularity conditions, where $D_0 = E\left[X_i X_i' \frac{1}{\pi(T_i,\gamma)} \Psi_\tau^2(\epsilon_i)\right]$, with $\epsilon_i = Y_i - X_i'\boldsymbol{\beta}(\tau)$ and $\Psi_\tau(t) = \tau - I(t < 0)$; $D_1 = E\Big[f_i(0|X_i) X_i X_i'\Big]$, with $f_i(t|X_i)$ denoting the conditional density function of ϵ_i given X_i; $I(\gamma) = E\Big[T_i T_i' \pi(T_i, \gamma)(1 - \pi(T_i, \gamma))\Big]$, $D_2 = E\Big[(1 - \pi(T_i, \gamma)) T_i X_i' \Psi_\tau(\epsilon_i)\Big]$. We assume that the symmetric matrices D_0, D_1 and $I(\gamma)$ are positive definite. As the matrix D_1 involves the unknown error density function, directly making inference about $\boldsymbol{\beta}(\tau)$ based on the asymptotic normal distribution is practically difficult. The results in Sherwood et al's paper were restricted to point estimator.

To construct a confidence interval for $\boldsymbol{\beta}(\tau)$ (abbreviated as $\boldsymbol{\beta}$ in the sequel when no confusion will be caused), we follow Yang, Wang and He (2015) and consider the asymmetric Laplace working likelihood

$$L(\boldsymbol{\beta}; D) = \frac{\tau^n (1-\tau)^n}{\sigma^n} \exp\Big(\frac{\sum_{i=1}^n \frac{R_i}{\pi(T_i,\gamma)} \rho_\tau(Y_i - X_i'\boldsymbol{\beta})}{\sigma}\Big), \tag{2}$$

where σ is a fixed scale parameter and γ is the parameter in the propensity score model. In practice, we replace the unknown nuisance parameter γ with a consistent estimator $\hat{\gamma}$. Given a prior $p_0(\boldsymbol{\beta})$, the posterior of $\boldsymbol{\beta}$ can be written as

$$p_n(\boldsymbol{\beta}; D) \approx p_0(\boldsymbol{\beta}) \exp\Big(\frac{\sum_{i=1}^n \frac{R_i}{\pi(T_i,\hat{\gamma})} \rho_\tau(Y_i - X_i'\boldsymbol{\beta})}{\sigma}\Big).$$

Following the same argument as in Yang, Wang and He (2015) and applying the derivations in Sherwood, Wang and Zhou (2013), we can show that

$$p_n(\boldsymbol{\beta}; D) \approx p_0(\boldsymbol{\beta}) \exp\Big(-\frac{n(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^W)^T D_1 (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^W) + o_p(1)}{2\sigma}\Big),$$

where $\hat{\boldsymbol{\beta}}_n^W$ is the weighted quantile regression estimator defined in (1). Our careful examination reveals that replacing the nuisance parameter γ with a reasonable estimator $\hat{\gamma}$ in constructing the working likelihood still leads to asymptotically valid inference on $\boldsymbol{\beta}$. Hence, with $p_0(\boldsymbol{\beta}) \approx 1$, the posterior of $\boldsymbol{\beta}$ is approximately normal with mean $\hat{\boldsymbol{\beta}}_n^W$ and variance $\hat{\Sigma} = \sigma D_1^{-1}/n$. Let $\hat{D}_0 = n^{-1}\sum_{i=1}^n \left[\frac{R_i}{\pi^2(T_i,\hat{\gamma})} X_i X_i' \Psi_\tau^2(Y_i - X_i'\hat{\boldsymbol{\beta}}_n^W)\right]$, $\hat{I} = n^{-1}\sum_{i=1}^n \left[T_i T_i' \pi(T_i,\hat{\gamma})(1-\pi(T_i,\hat{\gamma}))\right]$, $D_2 = n^{-1}\sum_{i=1}^n \left[\frac{R_i}{\pi(T_i,\hat{\gamma})}(1-\pi(T_i,\hat{\gamma}))T_i X_i' \Psi_\tau(Y_i - X_i'\hat{\boldsymbol{\beta}}_n^W)\right]$. The adjusted asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}_n^W$ is

$$\hat{\Sigma}_{\text{adj}} = \frac{n}{\sigma^2}\hat{\Sigma}\big(\hat{D}_0 - \hat{D}_2'\hat{I}^{-1}\hat{D}_2\big)\hat{\Sigma}. \tag{3}$$

We evaluate the above adjusted covariance matrix formula via a simple Monte Carlo experiment. We generate random data from the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 2$, $X_{i1} \sim U[0,1]$ and $X_{i2} \sim U[0,1]$. The covariate vector $X_i = (X_{i1}, X_{i2})'$ may be missing. Let R_i be an indicator variable which takes the value 1 if X_i is observed and 0 otherwise. We adopt logistic regression for the propensity score model $\log\big(P(R_i = 1|Y_i)/P(R_i = 0|Y_i)\big) = -1 + Y_i$, which results in about 20% missing covariates. We consider three different distributions for the random error ϵ_i: (1) N(0,1), (2) the heavy tailed distribution T_3, and (3) the heteroscedastic distribution $(1+0.5X_{i1})\xi_i$ where ξ_i's have the $N(0,1)$ distribution and are independent of the X_i's.

In the simulation experiment, we estimate the conditional median for the $N(0,1)$ and T_3 errors and estimate the 0.7 condtional quantile for the heteroscedastic error case. For sample sizes n=500 and 1000, we construct 90% and 95% confidence intervals for β_j, $j = 0, 1, 2$, using the adjusted covariance matrix formula in (3). Following Remark 1 in Yang, Wang and He (2015), we fix σ at a preestimated value, the inverse probability weighted maximum likelihood estimator of σ when the working likelihood in (2) is evaluated at the median. The empirical coverage probabilities of the confidence intervals are summarized in Table 1. We observe that the empirical coverage probabilities are close to the nominal levels, which confirms the effectiveness of the proposed approach for adjusting the covariance matrix estimation.

We would like to finish the discussion with suggestions on two possible future research directions. First of all, we believe it will be very interesting to consider an

Table 1: Empirical coverage probabilities of the confidence intervals for β_j ($j = 0, 1, 2$) for the missing data example

error distribution	nominal level	n	β_0	β_1	β_2
N(0,1)	0.90	500	0.882	0.902	0.891
N(0,1)	0.90	1000	0.891	0.892	0.901
N(0,1)	0.95	500	0.929	0.944	0.939
N(0,1)	0.95	1000	0.938	0.947	0.943
T_3	0.90	500	0.883	0.884	0.902
T_3	0.90	1000	0.898	0.893	0.911
T_3	0.95	500	0.934	0.940	0.952
T_3	0.95	1000	0.947	0.945	0.950
$(1 + 0.5X_1)\xi$	0.90	500	0.900	0.903	0.908
$(1 + 0.5X_1)\xi$	0.90	1000	0.889	0.904	0.904
$(1 + 0.5X_1)\xi$	0.95	500	0.943	0.952	0.948
$(1 + 0.5X_1)\xi$	0.95	1000	0.933	0.949	0.943

extension of the proposed approach to the more complex semiparametric setting. A representative example is partially linear quantile regression, which provides a useful model to accommodate nonlinearity and circumvent curse of dimensionality. If one is interested in making inference about the linear parameter (treating nonparametric component as nuisance), it seems that the proposed methodology is readily extendable. However, the asymptotic theory still requires careful derivation to incorporate the estimation of the nonlinear component. It is also curious to investigate if the proposed method can be useful in making inference on the nonparametric component. Second, an important practical advantage of the proposed approach is that it uses MCMC for computation. However, the examples of the paper have only studied models in rather low dimension. It will be interesting to further explore the scalability of the proposed approach to at least moderately larger dimension.

References

[1] Sherwood, B., Wang, L. and Zhou, A. (2013). Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in Medicine*, 32, 4967-4979.